DOI:
10.1002/anie.201704604

Document Version
Peer reviewed version

Simulated Isotope Exchange Patterns Enable Protein Structure Determination

Antoni J Borysik*[a]

Abstract: Understanding the myriad protein-protein interactions required for cell function requires efficient leveraging of biophysical data to drive computational docking. The detailed insight into protein interfaces provided by isotope exchange endows this experimental technique with a unique importance for docking approaches. However, progress in coupling these methods is hindered by the inability to interpret the complex exchange patterns in relation to protein structure. A method to simulate protein isotope exchange patterns from docking outputs is described and its utility to guide the selection of native assemblies demonstrated. Unique signatures are generated for each docking pose allowing high throughput ranking of whole docking simulations by pairwise comparison to experimental outputs. Native assemblies are obtained using nothing but their simulated profiles and experimental difference-data for individual proteins are sufficient to drive structure determination for the whole complex.

Cell function is orchestrated by a large network of protein-protein interactions (PPIs) which in humans is thought to involve around 500,000 binary contacts between proteins. [1] Protein docking has emerged to understand the abundance of PPIs at the molecular level in which the conformations of pairwise protein assemblies are predicted from the high resolution structures of their constituents. [2] Obtaining native structures from the high background of non-native state remains a significant challenge that can be met to some extent by utilizing available biophysical or biochemical information to drive the docking simulation. [3-5] Hydrogen deuterium exchange mass spectrometry (HDX-MS) can provide rapid and unique insight into protein-protein interfaces and has enormous potential for data-driven docking. This information is encoded by characteristic difference plots that report on local changes in the rate of heavy isotope uptake between monomeric and assembled states. [6-8] Unfortunately, our ability to efficiently leverage HDX-MS outputs for use with docking is limited as it currently impossible to evaluate the detailed isotope exchange patterns in relation to different protein structures. The development of methods that allow the generation of HDX-MS data directly from docking outputs would represent a significant advancement in the technique.

The potential for simulating HDX-MS data from protein atomic coordinates is an appealing strategy for structure determination. Central to this is the capacity to predict HDX protection factors (lnP) from an available structure which describe the degree to which the intrinsic isotope exchange rates deviate from those of a random coli in a folded protein. Unfortunately, our ability to accurately calculate lnP has proven problematic as the structural determinants of HDX remain unknown. This has resulted in the development of a range of different strategies that emphasize different aspects of protein structure from which lnP are derived. Solvent accessibility, hydrogen bonding, electrostatics and hydration have all been encoded by these methods to varying extents with values typically extracted from protein ensembles generated by molecular dynamics (MD) simulations or with advanced sampling techniques. [9-15] The ability of these methods to accurately predict lnP is however, generally low such that their capacity to select native structures from a high decoy background has never been previously reported. This has lead notable research to suggest that HDX phenomena are too complex to formulate *a priori* particularly where estimates are made from single conformations as would benefit structure determination. [16]

Here the modelling of a protein complex by HDX-MS is described and the utility of the approach demonstrated with the binary assembly formed between the enzyme α-chymotrypsin and the proteinase inhibitor eglin c. [10] The approach involves the simulation of HDX-MS data directly from protein docking outputs permitting the critical evaluation of individual assemblies and high throughput quantitative ranking of the entire docking simulation. Native structures (iRMSD < 0.7 Å) are extracted using nothing but their simulated HDX profiles as restraints and experimental difference plots for either protein in the assemblies are shown to be sufficient to guide structure determination of the entire complex.

HDX-MS experiments were performed on a Synapt G2Si HDMS coupled to an Acquity UPLC M-Class system with HDX and automation (Waters Corporation, Manchester UK). The isotope uptake of eglin c and α-chymotrypsin was determined by continuous labelling on individual and premixed samples and difference plots prepared by subtraction of the bound from the unbound HDX-MS patterns. Protein docking was initially performed with PatchDock resulting in ~ 5800 transformations which were then refined with FiberDock to remove steric overlap and allow for flexibility. [17] HDX-MS data were simulated by estimating the protection factors (lnP) of backbone amide protons from protein structures according to previous descriptions of equilibrium exchange (Equation 1). [18] In Equation 1 the protection of residue *i* is expressed as the number of heavy atoms (N_i^C) and hydrogen bond acceptors (N_i^H) within defined distance cut-offs of the backbone amide each weighted by an empirically determined scaling term (β). Although several different formulations of lnP have been proposed this approach benefits from ease of implementation enabling high throughput ranking of an entire docking simulation (Supporting Information).

(Equation 1) $$lnP_i^{sim} = N_i^C \beta_C + N_i^H \beta_H$$

lnP were calculated directly from each structure in the docked assemblies such that atoms from neighboring chains within the distance cut-offs contributed to the lnP of exchangeable residues. The simulated lnP were then used to modify the intrinsic exchange rates (k_{int}) and HDX-MS outputs prepared for each protein in the assemblies in the form of difference data between the bound and unbound states. The simulations captured the overall features of the experimental data requiring only moderate optimization of the output amplitudes. Additional calculations were made on each docking pose to extract lnP overlooking any contribution from the neighboring chain. This allowed the definition of two different monomeric states both prior to, and following protein docking, the structures of which typically varied in each docking pose due to flexible refinement. Accordingly two HDX-MS difference plots were simulated for each protein in an assembly that either included (flexible) or ignored (rigid) structural changes during docking (Supporting Information, Figure 1).

[a] Dr, AJ, Borysik
Department of Chemistry
King's College London
Britannia House
London, SE1 1DB
E-mail: Antoni.borysik@kcl.ac.uk

Supporting information for this article is given via a link at the end of the document.

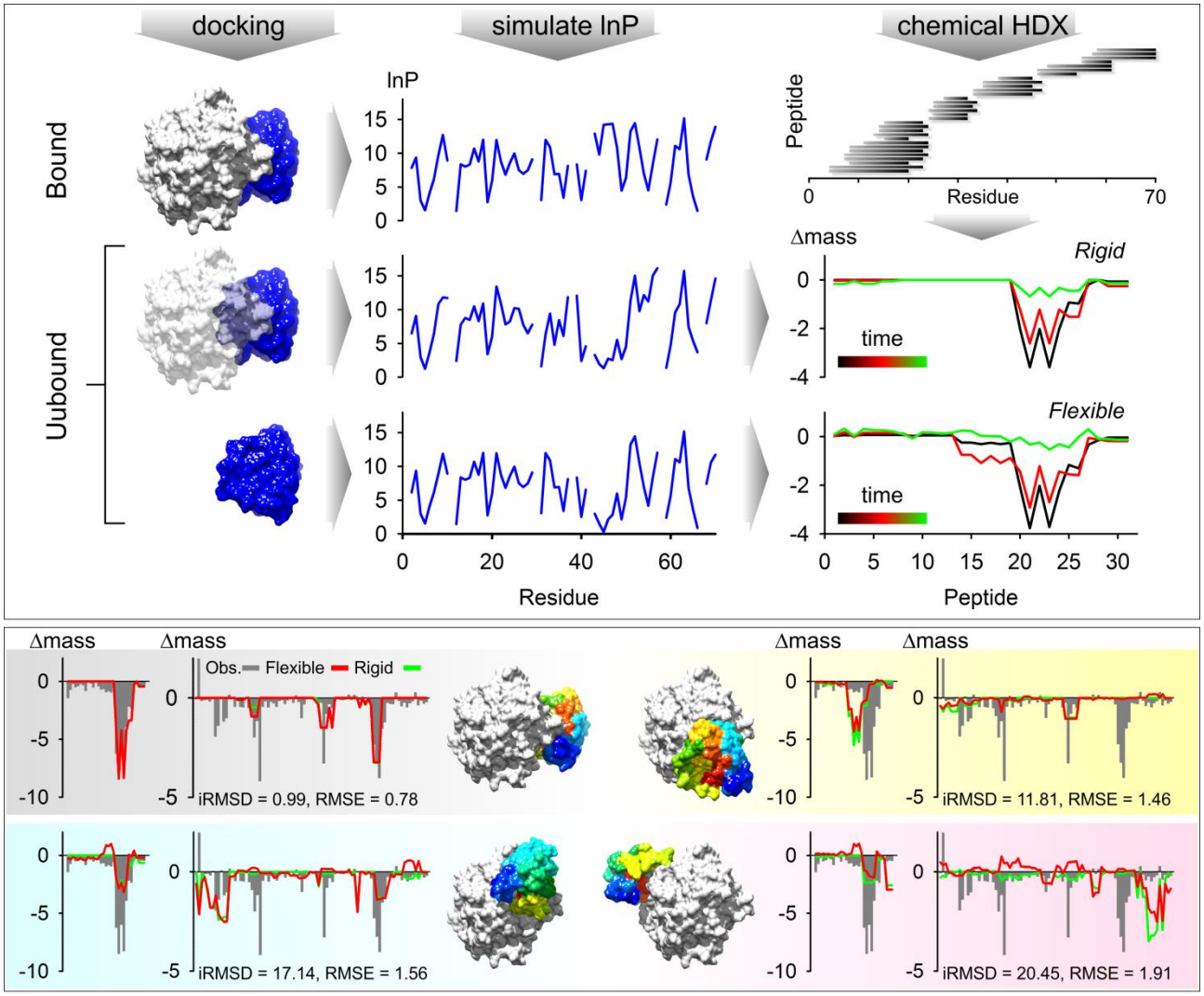


Figure 1: Workflow for HDX-MS data simulation and analysis. (Upper panel) eglin c (blue) and α-chymotrypsin (gray) were docked and lnP simulated for each protein. The intrinsic exchange rates (k_{int}) were determined and then modified by the lnP to generate HDX-MS data for each protein mapped onto the experimental peptide list. Subtraction of the bound from the unbound HDX-MS data generated two difference plots for each protein which either considered or ignored flexible fitting depending on the unbound dataset used. The figure shows example outputs for eglin c only at three labelling time points. (Lower panel) Experimental (gray) and simulated (red and green) HDX-MS difference plots for different docking poses. The top structure ranked by RMSE is shown (gray background) along with three alternate docking poses. The iRMSD with the native state (1ACB) is shown along with the average RMSE of the simulated HDX-MS data. In all cases the left hand plots are for eglin c and the Δmass reports the summed values over three labelling time points.

The utility of the approach to guide selection of native structures from the docking simulation was then evaluated. The experimental HDX-MS difference plots of each protein were compared with the simulated profiles of the corresponding protein in each pose allowing the definition of an RMSE-based metric and quantitative scoring of each structure. For non-native states the RMSE of the simulated HDX-MS profiles correlate poorly with the interfacial RMSD (iRMSD) of the corresponding structures due to the range of possible orientations that can generate equally poor RMSE scores. However, as the iRMSD decreases the simulated HDX-MS profiles align more closely with the experimental difference plots resulting in higher convergence between the RMSE and iRMSD values (Figure 2 a – c). Overall the HDX-MS simulations of eglin c outperform those of α-chymotrypsin with the top 1% ranked assemblies (58 poses) having respective mean alignment scores with the native structure of 2.1 Å and 4.4 Å iRMSD. Combining the RMSE scores of both proteins moderately improves the ability to identify native structures with the top 1% docking outputs having an average alignment of 2.0 Å iRMSD with the native complex. The top 1% poses ranked by the Eglin c simulations contain relatively few decoys with 85% of these structures also in the top 1% when the assemblies are ranked according to their iRMSD. This number reduces to 60% for α-chymotrypsin presumably reflecting the overall weaker fit of the HDX-MS simulations of this protein. Nevertheless, all of the top 1% docking poses ranked by the HDX-MS simulations of α-chymotrypsin successfully place the inhibitor protein in the correct binding interface. This suggests a degree of tolerance in the approach that allows useful insight into correct binding modes even when native state simulations only weakly accommodate the experimental data.

The top ranking assemblies identified by the HDX-MS simulations of either protein are sufficiently enriched with native poses to allow the identification of these species. The top 1% poses ranked by the HDX-MS profiles were clustered following an all-vs-all alignment. Docking poses belonging to the major subtree < 1.0 Å pairwise distance were obtained and in all cases these assemblies differed by no more than 0.7 Å iRMSD with the native assembly (Figure 2 d – f). The HDX-MS simulations

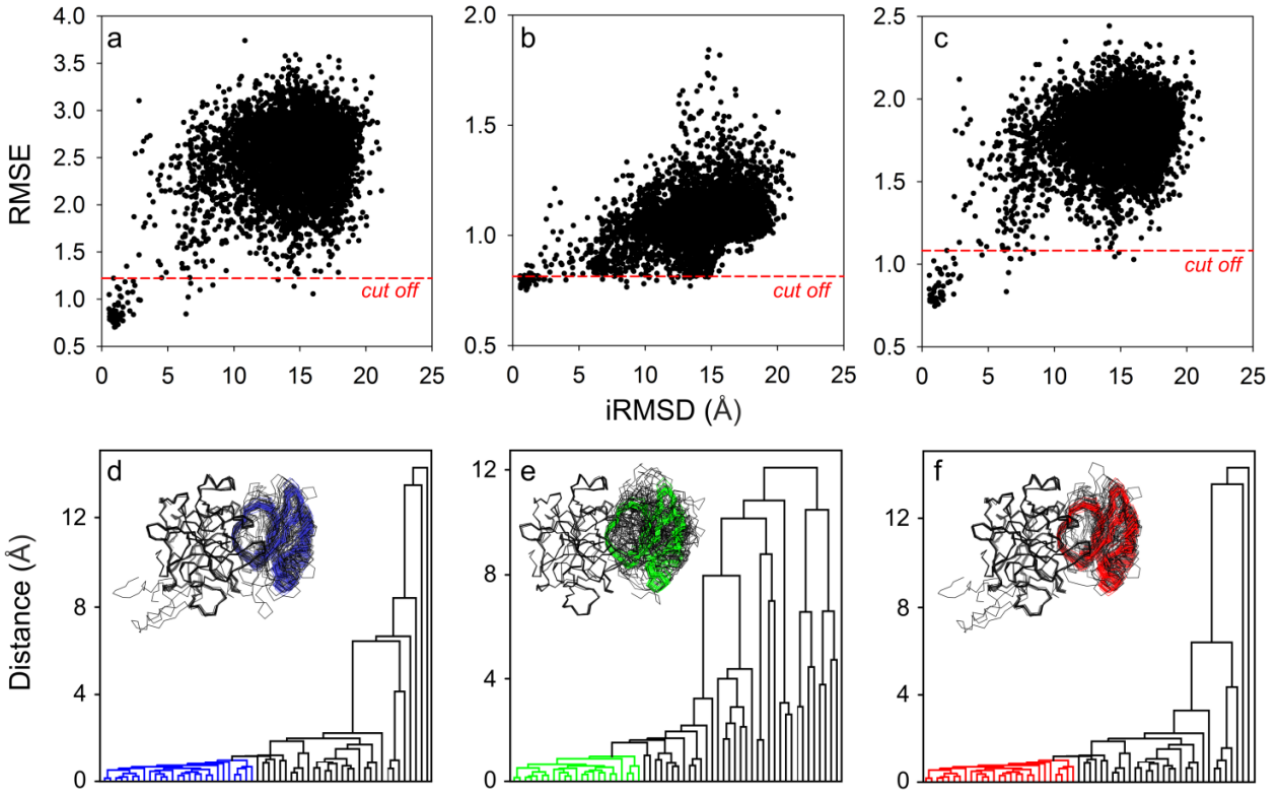


Figure 2: Evaluation of HDX-MS-driven protein docking. (a – c) relationship between the iRMSD and RMSE of the simulated and experimental HDX-MS difference plots for ~ 5800 docking poses. The RMSE was determined by comparing the simulated and experimental HDX-MS difference data for (a) eglin c (b) α-chymotrypsin or (c) the binary assembly. The cut-off for poses with the top ranked 1% RMSE is indicated in each plot. (d – e) dendrograms for the all-vs-all clustering outputs of the top 1% RMSE ranking structures for (d) eglin c (e) α-chymotrypsin or (f) the binary assembly. Insert shows the structural alignment of the top 1% RMSE structures (58 docking poses) with those belonging to the major subtree < 1.0 Å distance colored as shown in the associated dendrograms

provide a set of unique signatures for each docking output from which native states can be identified. Surprisingly, experimental HDX-MS difference data for either protein alone are shown to be sufficient to successfully guide structure determination for the whole complex.

The identification of native poses from a high background of decoys is a major challenge in protein docking. Information-driven methods address this issue by utilizing experimental restraints to better define the overall search space and reduce the range of outputted decoys. [19, 20] HDX-MS difference plots allow the identification of residues residing in protein-protein interfaces which can then be activated prior to protein docking. However, this binary implementation of HDX-MS restraints only allows residues to be switched into on/off positions and overlooks the overall shape of the experimental difference profile. Post-docking evaluation remains a significant challenge and discriminating between different poses by HDX-MS is still largely a qualitative procedure.

Unlocking the potential of HDX-MS for structure determination requires the ability to simulate experimental data directly from high resolution modelling outputs. However, these calculations are ambiguous because the aspects of protein structure that protect exchangeable sites from HDX have yet to be properly defined. HDX is also an ensemble property and no single state should therefore fully describe the exchange behavior of a protein. The implicit encoding of protein ensembles by molecular dynamics (MD) simulations can provide better representations of the structures responsible for HDX. [21] However, the benefits of extracting lnP from MD ensembles is usually marginal probably reflecting limitations in current lnP formulations along with fundamental differences between MD and HDX with regard to their timescales and amplitudes of motion. A requirement for computationally expensive processing also imposes significant barriers on throughput such that the evaluation of entire docking outputs would not be feasible.

In the present work the ability to identify native structures using HDX-MS patterns simulated directly from atomic coordinates is demonstrated. This is facilitated by focusing on the generation of difference patterns for each protein which localises the simulated outputs to the interacting regions. Accurate projections of isotope uptake across the entire protein backbone are not required and the difference patterns naturally bias the calculations towards the native pose. Dynamical changes on binding can also be considered implicitly by incorporating structural changes brought about by flexible refinement into the difference plot calculations. Data generation can be achieved with high throughput and post-docking evaluations performed quantitatively and with improved scrutiny.

The current expression used to predict lnP from protein structures was initially trained on a limited dataset of globular proteins and is not tuned for protein interfaces. Unsurprisingly, it required prior optimization and will require further characterisation against a larger dataset of previously determined PPIs before it can be used with ab initio applications. Nevertheless, the present work clearly demonstrates that complex HDX phenomena can be captured by a simple expression to sufficient quality to guide the selection of native structures from a high background of non-native states. The integration of HDX-based scores with current scoring functions should result in improved functionality of this method. Extension of the present approach to characterise different docking systems such as those involving drugs, peptides or lipids will be most interesting.

Acknowledgements

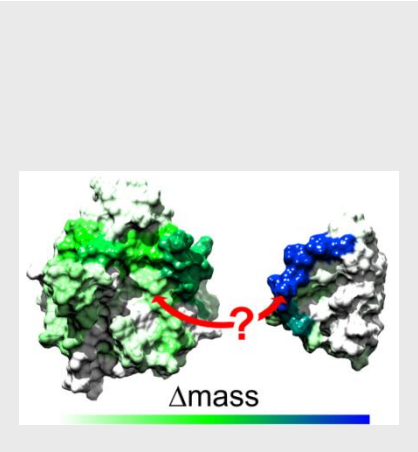
The author acknowledges financial support from the Royal Society sponsor reference RG150222.

Keywords: Protein-protein interactions • Protein structures • Mass spectrometry • Isotopic labelling

1. Mitra, K., et al., *Nat Rev Genet*, 2013. **14** 719-732.
2. Keskin, O., N. Tuncbag, and A. Gursoy, *Chem Rev*, 2016. **116** 4884-4909.
3. Gromiha, M.M., K. Yugandhar, and S. Jemimah, *Curr Opin Struct Biol*, 2016. **44** 31-38.
4. Moreira, I.S., et al., *Phys Chem Chem Phys*, 2015. **17** 2378-2387.
5. Schneidman-Duhovny, D., M. Hammel, and A. Sali, *J Struct Biol*, 2011. **173** 461-471.
6. Anand, G.S., et al., *Proc Natl Acad Sci U S A*, 2003. **100** 13264-13269.
7. Zhang, Z. and D.L. Smith, *Protein Sci*, 1993. **2** 522-531.
8. Houde, D., S.A. Berkowitz, and J.R. Engen, *J Pharm Sci*, 2011. **100** 2071-2086.
9. Vendruscolo, M., et al., *J Am Chem Soc*, 2003. **125** 15686-15867.
10. LeMaster, D.M., J.S. Anderson, and G. Hernandez, *Biochemistry*, 2009. **48** 9256-9265.
11. Craig, P.O., et al., *J Am Chem Soc*, 2011. **133** 17463-17472.
12. Liu, T., et al., *J Am Soc Mass Spectrom*, 2012. **23** 43-56.
13. Zhang, Y., et al., *Biochemistry*, 2014. **53** 5619-5630.
14. Park, I.H., et al., *J Chem Inf Model*, 2015. **55** 1914-1925.
15. Devaurs, D., et al., *Front Mol Biosci*, 2017. **4** 1-13.
16. Skinner, J.J., et al., *Protein Sci*, 2012. **21** 987-995.
17. Mashiach, E., R. Nussinov, and H.J. Wolfson, *Proteins*, 2010. **78** 1503-1519.
18. Best, R.B. and M. Vendruscolo, *Structure*, 2006. **14** 97-106.
19. Karaca, E., et al., *Mol Cell Proteomics*, 2010. **9** 1784-1794.
20. Rey, M., et al., *Structure*, 2014. **22** 1538-1548.
21. Radou, G., et al., *Biophys J*, 2014. **107** 983-990.

COMMUNICATION

Determining protein structures by isotope exchange. Protein complexes are assembled by simulating their complex isotope patterns and comparing these outputs to those obtained by experiment.



*Antoni J Borysik**

Page No. – Page No.

Simulated Isotope Exchange Patterns Enable Protein Structure Determination